Spotting the target: microarrays for disease gene discovery

Paul S Meltzer

Microarray technologies enable genome-scale expression measurements. Already proved to be of value for the functional analysis of individual genes and biological processes, the application of expression profiling to disease gene discovery is now growing in importance and practicality.

Addresses
Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892-4470, USA; e-mail: pmeltzer@nhgri.nih.gov

Current Opinion in Genetics & Development 2001, **11**:258–263

0959-437X/01/$ – see front matter

Abbreviations

CGH	comparative genomic hybridization
NF2	neurofibromatosis type 2
PARP	poly(ADP-ribose) polymerase
TD	Tangier disease

Introduction

Positional cloning projects have been greatly facilitated by the availability of increasingly precise maps and sequence databases for diverse species. This same avalanche of genomic data has inspired an intense effort to study aspects of genome function in a high-throughput fashion. The parallel analysis of gene expression has emerged as one of the most productive embodiments of this approach.

Practical technologies for large-scale gene-expression analysis are now being widely implemented. Microarrays comprising either oligonucleotides or cDNA fragments representing thousands of genes are well suited to the analysis of multiple samples [1,2]. To obtain genome-scale expression data, mRNA from the source of interest is converted to an appropriately labeled form and hybridized to the microarray. Both radioactive and fluorescence-detection strategies are in use to measure the resulting hybridization signal. The resulting raw data — an image obtained from a fluorescence scanner or phorphorimager — is processed with computer software to generate a spreadsheet of gene-expression values. The application of statistical techniques to microarray data allows classification and class discovery within a group of samples, and clustering of genes according to their pattern of expression.

Microarrays have been successfully applied to characterize biological processes and to dissect pathways downstream of a particular gene of interest. Studies in the yeast *Saccharomyces cerevesiae*, with its relatively small genome and highly tractable genetics, have led the way and continue with recent reports on signal transduction [3], meiosis [4] and transcript localization [5]. Despite the challenges posed by their genome sizes, large-scale expression analysis in mammals is also becoming increasingly productive.

As the technology for microarray analysis has matured and disseminated, new applications continue to be developed. One frequently discussed area is the potential use of microarray expression analysis in projects to positionally clone and discover disease genes. Although reviews of this topic outnumber reports of concrete achievement, it is appropriate to examine the state of the art and to consider how microarray analysis might accelerate these types of research. I discuss these points, together with recent developments in microarray research, in this review.

How might microarrays help find hereditary disease genes?

Several major approaches to locating hereditary disease genes might be imagined. In the simplest case, the target gene of interest might be identified directly by characteristic changes in expression level across a series of samples. Alternatively, statistical analysis of microarray data might aid gene discovery by revealing pathways related to the target gene and facilitating identification of candidate genes.

Microarrays can also be used to analyze genomic DNA rather than mRNA. This is illustrated by the special case of copy-number change in cancer, where it is possible to use array-format comparative genomic hybridization (CGH) to define genes associated with cancer progression [6••,7,8•]. In CGH, gene copy number is measured in a DNA sample labeled with one fluorochrome by comparison to the signal obtained by simultaneous hybridization of normal DNA labeled with a second fluorochrome. In principle, copy-number data can be linked to expression data to define a list of candidate target genes associated with gain of chromosomal regions [9•,10]. Although there is no example to date, tumor suppressors might be mapped by linking loss of gene expression to regions of deletion in tumors.

Of course, microarrays can be used as sophisticated dot blots to screen arrays of clones isolated with techniques such as RDA [11]. (RDA [representational difference analysis] is a PCR-based subtraction technique that can be used to isolate DNA fragments that vary in abundance between two sources.) Stephan *et al.* [12] have identified exons of the Niemann–Pick Type C disease isolated from arrayed genomic sequences using mRNA from cells differentially expressing *NPC1*. Finally, genes might be linked to specific phenotypes, particularly in yeast, through methods that allow genome-wide mutational screens using microarrays as a readout [13].

Finding the best candidate

It is enticing to hope that analysis of microarray data might lead to the direct identification of disease genes. Ideally, one would compare a group of samples of varying genotype and identify good candidate genes by their pattern of gene

expression. The expected signature of a mutant gene is reduced expression level in samples with the abnormal allele. For this strategy to work, the mutant allele would have to be either deleted or result in a poorly expressed transcript.

Fortunately, the phenomenon of nonsense-mediated decay of mRNA gives some reason to hope that this result might actually be achieved. Nonsense-mediated decay (reviewed in [14]) results in the degradation of certain mRNAs containing premature termination codons. This phenomenon has been observed in a number of disease genes [15,16].

In addition, abnormalities in 3′-untranslated region structure that interfere with normal polyadenylation may also lead to reduced survival of transcripts [17]. A reduction in steady-state mRNA levels of disease genes cannot be assumed, however, because the competence of a transcript to undergo nonsense-mediated decay is variable and some mutations may result in exon skipping [18,19], as has been shown by Liu *et al.* [20] for the *BRCA1* gene. This strategy also requires a sufficient number of samples from cells or tissues affected by the disease to help optimize the downstream data analysis.

Although obvious, an additional requirement of expression-based strategies is that the target gene is actually represented on the microarrays used. Although arrays of more than 10,000 genes are commonplace and complete genome microarrays can be anticipated, they are not yet routinely available. It is also probably unrealistic to assume that only a single gene or a few genes will stand out from the crowd with sufficient clarity to allow easy candidate selection. More likely, a strategy combining positional information with expression information will be necessary.

This combination of approaches has been used by Lawn *et al.* [21••] in the discovery of the Tangier disease (TD) gene *ABC1*. Microarray analysis led to the generation of a list of 175 cDNAs underexpressed by 2.5-fold or more in the fibroblasts of an affected individual. By combining this data with linkage information that localized the disease gene to chromosome 9q between the markers WI-14706 and WI-4062, the candidate list was narrowed sufficiently to identify the gene *ABC1*, which did indeed carry mutations.

Notably, Lawn *et al.* [21••] used commercial cDNA arrays containing 58,800 cDNAs, which presumably provided a reasonably thorough genome scan. One might imagine that regional searches could be made by constructing targeted microarrays covering a particular candidate region. This has been done for the X chromosome and for chromosome 17q [9•,22••].

It is important to bear in mind that almost all research employing microarray expression analysis depends heavily on statistical analysis to extract the most useful information from the huge number of data points generated. This means that any investigator attempting to use microarrays for disease gene discovery will also seek to go beyond this direct type of search and also examine the broader effects of mutation on gene expression in samples from affected individuals.

If one were not able to identify easily a candidate gene by virtue of its underexpression, perhaps the recognition of pathways altered consistently across a set of specimens might lead to the identification of good candidate genes or, at the very least, might illuminate some aspects of pathogenesis.

Finding the disease pathway affected by known genes

The complexity of microarray data is illustrated by another interesting feature of the TD data — the overexpression of 375 cDNAs by 2.5-fold or more. This result, revealing a total of 550 cDNAs with altered expression, is probably typical of what might be expected in most projects. In addition to innumerable technical factors, variations in gene expression across samples might be due to random fluctuations or confuting variables such as age, sex, site of sample and irrelevant genetic variations. Still, it would seem reasonable to suppose that the presence of a mutation in a pathway might frequently lead to secondary events affecting the level of expression of many other genes functionally connected to the disease gene.

Most published examples attempting to place genes from microarray data on samples carrying mutations into coherent pathways are in the setting of model systems for which the mutation is already known. McNeish *et al.* [23•] have examined a mouse model of TD with microarrays containing 11,000 genes and have identified 131 genes with greater than 1.8-fold differential regulation, many of which can be grouped into a few function-related categories. Their study demonstrates how studies of a relatively tractable experimental model can enhance the value of data obtained from human samples.

Likewise, Soukas *et al.* [24•] examined gene expression in white adipose tissue from mice expressing varying levels of the leptin gene. Seventy-seven genes were dysregulated by threefold or more in these *ob/ob* mice, including a number of key genes in fat metabolism. One cluster of genes was coordinately regulated by SREBP-1/ADD1, but the regulating mechanisms linking genes in several other clusters remain unknown. Although the complete pattern of changes observed cannot be explained as yet, the relevance of the leptin gene to fat metabolism is amply demonstrated.

Simbulan-Rosenthal *et al.* [25•] examined fibroblasts from mice deficient in poly(ADP-ribose) polymerase (PARP) with microarrays covering 11,000 genes and identified 91 genes differentially regulated by at least twofold relative to wild-type fibroblasts. About 40% of these could be related to either the cell cycle or remodeling of the cytoskeleton or extracellular matrix — processes known to be associated with PARP function.

Callow *et al.* [26] examined livers from apolipoprotein AI knockout mice, scavenger receptor B1 transgenic mice and wild-type mice on microarrays containing 5600 cDNAs. They used *t*-test statistics to identify a small number of genes that differed significantly across these conditions.

For disease gene discovery, the interpretation of expression data in terms of pathways is more difficult because there is no *a priori* knowledge of the disease gene function. This leads to a consideration of the process of grouping differentially expressed genes into pathways.

Placing genes in pathways to gain clues about unknown genes

Can pathways actually be discerned from microarray data? It is worthwhile considering some of the individual steps in the process of deducing pathway information from these data. Clustering of genes into co-regulated groups is computationally straightforward and readily generates this type of information [27]. Similarly, there has been great success in classifying biological samples from microarray data, particularly for cancer specimens [28•,29••–31••,32]. These studies are promising in identifying critical genes for cancer progression at the expression level, although these are not necessarily 'disease genes' in the genetic sense [33••].

Nonetheless, Hedenfalk *et al.* [34••] have even shown that it is possible to sort breast cancer specimens according to the presence of hereditary mutations in *BRCA1* or *BRCA2*. One of the most striking results in their study was the demonstration that a sample that clustered with those from patients carrying mutations in *BRCA1* lacked a *BRCA1* mutation but was highly methylated at the *BRCA1* promoter.

It might be hoped that this approach could aid complex disease gene discovery by sorting samples into groups that share a common genetic defect. When combined with positional data from linkage analysis, such an approach might be expected to take on a significant role in the study of complex disease.

In contrast to clustering samples and genes, the interpretation of expression data to infer the pathway affected by a disease gene mutation is much more problematic. The initial problem one faces in this type of analysis is the limited annotation of the genome. When examining an expression database, one immediately encounters difficulty in placing genes into functional categories. This is beset with a number of obstacles, the first of which are the numerous aliases that confuse gene nomenclature.

The introduction of two on-line resources, LocusLink and Refseq, have gone a long way towards overcoming this problem by providing a unique identifier and curated sequence for each gene [35]. This is absolutely critical to the next phase of analysis, which is the cross-reference to other databases of gene function including, most importantly, literature databases. Frequently, different functions or interpretations of gene function are linked to distinct aliases for a given gene. Only by thoroughly combing the literature, can the most comprehensive picture of gene function be obtained. Substantial efforts are being made to organize the genes of known function into meaningful categories.

Although a detailed discussion of the problem of gene annotation is beyond the scope of this review, the public availability of certain resources should be noted. In particular, the Gene Ontology consortium uses a common language to organize functional information in all species [36]. Currently, the Gene Ontology database contains database links for *Drosophila*, *S. cerevesiae*, mouse and *Caenorhabditis elegans*. Genes are categorized in three hierarchical schemes according to molecular function, biological process and cellular component.

Methods to process groups of genes with respect to literature databases are also under development [37–39]. One system, High-density Array Pattern Interpreter (HAPI; http://array.ucsd.edu/hapi/), is publicly available. It is anticipated that search engines that can carry out these computations with the output of expression databases will significantly accelerate the process of organizing data from microarrays.

Although it is relatively straightforward to identify lists of genes that are co-regulated across a set of samples, this may not be a sufficiently sensitive method to extract functionally related genes. Intensive efforts to establish alternate computational methods are continuing.

Seungchan *et al.* [40] have described a multivariate technique that has the potential to identify relationships among genes that are refractory to methods based on linear correlation. Akutsu *et al.* [41] have proposed a method for modeling gene expression in terms of Boolean networks, whereas Friedman *et al.* [42] have proposed a Bayesian method. Hastie *et al.* [43] have described a method termed 'gene shaving', which differs from hierarchical clustering in that genes may belong to more than one cluster. Brown *et al.* [44] have advocated the use of method based on the theory of 'support vector machines', a computer learning method that they have adapted to the functional categorization of expression data.

Finding regulatory motifs

One great challenge remaining in the analysis of mammalian expression data will be to link this information to regulatory elements in the genome sequence. Promising results in yeast continue to appear. Iyer *et al.* [45••] have taken advantage of the small size of the yeast genome to array non-coding DNA and identify the genes regulated by the cell-cycle transcription factors SBF and MBF. Ren *et al.* [46••] have achieved similar results for Gal4 and Ste12, and Livesey *et al.* [47] have identified the response element configuration and genes responsive to the mouse homeobox gene *Crx*.

The development of progressively more sophisticated computational methods increases optimism that genes related to a phenotype can be accurately extracted and placed in functionally related groups to help generate new hypotheses. Even with this goal accomplished, one would expect that the effects of mutation on one biochemical pathway will radiate to affect numerous other pathways. Identifying the pathway primarily affected will be a significant challenge.

Using microarrays to map genomic DNA

Although using microarrays to identify regions of copy-number change in cancers has received the most attention, array format CGH might also be useful for mapping hereditary disease genes. Bruder *et al.* [48••] have used microarrays tiled across a 7-Mb region including the neurofibromatosis type 2 gene (NF2) to analyze DNA from 116 NF2 patients. Using this exquisitely accurate system, they were able to identify 24 patients with gene deletions and show that there was no correlation with disease severity.

In principle, this type of approach could be applied to a region containing an unknown disease gene. Because positional cloners frequently assemble contigs covering regions of linkage, the availability of genomic clones may not be problematic. However, the technology for arraying and accurately determining copy number in this setting is still confined to a few laboratories.

Conclusions

Unquestionably, large-scale expression analysis is now established in the study of genome function. The power of this approach continues to be enhanced by technical advances and, importantly, by the development of very large coherent expression databases from samples collected across a broad range of conditions [49••]. The recent report from Shoemaker *et al.* [50] points to the future with microarrays composed of over one million oligonucleotides representing 442,785 exons predicted from the draft human genome sequence. These developments suggest that microarray analysis will increasingly merit consideration as an ancillary technique to facilitate hereditary disease gene discovery.

Update

Loftus and Pavan have recently used melanocyte-specific microarrays to identify a mouse coat color gene (S Loftus, W Pavan, personal communication).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467-470.

2. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP: **Accessing genetic information with high-density DNA arrays.** *Science* 1996, **274**:610-614.

3. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR *et al.*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.** *Science* 2000, **287**:873-880.

4. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE: **The core meiotic transcriptome in budding yeasts.** *Nat Genet* 2000, **26**:415-423.

5. Takizawa PA, DeRisi JL, Wilhelm JE, Vale RD: **Plasma membrane compartmentalization in yeast by messenger RNA transport and a septin diffusion barrier.** *Science* 2000, **290**:341-344.

6. •• Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41-46.

This study shows that cDNA microarrays can be used to determine gene copy number by comparative genomic hybridization. Using the same array for copy number and expression measurements enables rapid mapping of regions of gene amplification in cancers. It also raises the possibility for finding mutational targets associated with copy-number loss and reduced expression.

7. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y *et al.*: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nat Genet* 1998, **20**:207-211.

8. • Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ: **Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays.** *Genome Res* 2000, **10**:1126-1137.

An oligonucleotide microarray designed to cover 600 SNPs yields allelic imbalance data on about 150 markers in analyses of tumor-cell DNA. However, the method cannot distinguish between loss and gain at a locus.

9. • Barlund M, Monni O, Kononen J, Cornelison R, Torhorst J, Sauter G, Kallioniemi O-P, Kallioniemi A: **Multiple genes at 17q23 undergo amplification and overexpression in breast cancer.** *Cancer Res* 2000, **60**:5340-5344.

This study shows the potential of tissue microarrays to establish rapidly the rate of gene amplification at a series of loci on 17q23 in 372 breast cancers. This is an excellent illustration of the complementary application of tissue and expression microarrays.

10. Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle AA, Krahe R: **Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics.** *Proc Natl Acad Sci USA* 2001, **98**:1124-1129.

11. Welford SM, Gregg J, Chen E, Garrison D, Sorensen PH, Denny CT, Nelson SF: **Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization.** *Nucleic Acids Res* 1998, **26**:3059-3065.

12. Stephan DA, Chen Y, Jiang Y, Malechek L, Gu JZ, Robbins CM, Bittner ML, Morris JA, Carstea E, Meltzer PS *et al.*: **Positional cloning utilizing genomic DNA microarrays: the Niemann-Pick type C gene as a model system.** *Mol Genet Metab* 2000, **70**:10-18.

13. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al.*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901-906.

14. Frischmeyer PA, Dietz HC: **Nonsense-mediated mRNA decay in health and disease.** *Hum Mol Genet* 1999, **8**:1893-1900.

15. Christiano AM, Amano S, Eichenfield LF, Burgeson RE, Uitto J: **Premature termination codon mutations in the type VII collagen gene in recessive dystrophic epidermolysis bullosa result in nonsense-mediated mRNA decay and absence of functional protein.** *J Invest Dermatol* 1997, **109**:390-394.

16. Jacoby LB, MacCollin M, Parry DM, Kluwe L, Lynch J, Jones D, Gusella JF: **Allelic expression of the NF2 gene in neurofibromatosis 2 and schwannomatosis.** *Neurogenetics* 1999, **2**:101-108.

17. Muhlrad D, Parker R: **Aberrant mRNAs with extended 3′ UTRs are substrates for rapid degradation by mRNA surveillance.** *RNA* 1999, **5**:1299-1307.

18. Bauer JW, Rouan F, Kofler B, Rezniczek GA, Kornacker I, Muss W, Hametner R, Klausegger A, Huber A, Pohla-Gubo G *et al.*:

A compound heterozygous one amino-acid insertion/nonsense mutation in the plectin gene causes epidermolysis bullosa simplex with plectin deficiency. *Am J Pathol* 2001, **158**:617-625.

19. Romao L, Inacio A, Santos S, Avila M, Faustino P, Pacheco P, Lavinha J: **Nonsense mutations in the human β-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation.** *Blood* 2000, **96**:2895-2901.

20. Liu HX, Cartegni L, Zhang MQ, Krainer AR: **A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.** *Nat Genet* 2001, **27**:55-58.

21. •• Lawn RM, Wade DP, Garvin MR, Wang X, Schwartz K, Porter JG, Seilhamer JJ, Vaughan AM, Oram JF: **The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway.** *J Clin Invest* 1999, **104**:R25-R31.
This study illustrates the potential of using gene-expression analysis in concert with positional information to identify a hereditary disease gene.

22. •• Sudbrak R, Wieczorek G, Nuber UA, Mann W, Kirchner R, Erdogan F, Brown CJ, Wohrle D, Sterk P, Kalscheuer VM *et al.*: **X chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications.** *Hum Mol Genet* 2001, **10**:77-83.
The authors construct a specialized microarray containing 2423 cDNAs from the X chromosome. They confirm its utility with tests on samples containing varying numbers of X chromosomes. Notably, they can identify three genes contained within a male-viable deletion.

23. • McNeish J, Aiello RJ, Guyot D, Turi T, Gabel C, Aldinger C, Hoppe KL, Roach ML, Royer LJ, de Wet J *et al.*: **High density lipoprotein deficiency and foam cell accumulation in mice with targeted disruption of ATP-binding cassette transporter-1.** *Proc Natl Acad Sci USA* 2000, **97**:4245-4250.
The authors use gene-expression analysis in a mouse knockout model of Tangier disease. This study provides an example of the progress and problems in identifying pathways downstream of disease genes. It also illustrates the advantages of studying a disease model in the mouse for which target tissues can be readily obtained.

24. • Soukas A, Cohen P, Socci ND, Friedman JM: **Leptin-specific patterns of gene expression in white adipose tissue.** *Genes Dev* 2000, **14**:963-980.
This study demonstrates that a large proportion of the genes altered in mice carrying mutated leptin genes can indeed be placed in pathways credibly related to leptin function.

25. • Simbulan-Rosenthal CM, Ly DH, Rosenthal DS, Konopka G, Luo R, Wang ZQ, Schultz PG, Smulson ME: **Misregulation of gene expression in primary fibroblasts lacking poly(ADP-ribose) polymerase.** *Proc Natl Acad Sci USA* 2000, **97**:11274-11279.
This study explores the altered patterns of gene expression in mice deficient for poly(ADP-ribose) polymerase, and links the observed changes to pathways affecting cell cycle, DNA replication, the extracellular matrix and the cytoskeleton.

26. Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice.** *Genome Res* 2000, **10**:2022-2029.

27. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.

28. • Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**:227-235.
The authors show that expression profiles determined by cDNA microarrays analysis can be used to cluster a variety of cancer cell lines into distinct disease categories.

29. •• Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503-511.
By using a cDNA microarray (the 'lymphochip') optimized for relevant gene content, the authors are able to subclassify diffuse large B-cell lymphoma specimens into two groups. Of importance, they can relate these groups to normal B-cell differentiation and show that the group of patients with a germinal-center-like phenotype has a better prognosis.

30. •• Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A *et al.*: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406**:536-540.
In this study, the potential of microarrays for class discovery is demonstrated by the identification of an unknown subset of melanoma samples with a characteristic gene-expression profile. By comparing this profile with that of an *in vitro* model of melanoma metastasis, the authors can predict biological properties (such as reduced motility) associated with this subset, which they confirm experimentally.

31. •• Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747-752.
The authors use cDNA microarrays to establish the gene-expression profile of breast cancer specimens. The results are striking and suggest distinct histogenesis for subsets of breast cancers. Along with Bittner *et al.* [30••], this study shows that two samples from the same patient tend to a greater similarity than any pair within a sample set.

32. Amler LC, Agus DB, LeDuc C, Sapinoso ML, Fox WD, Kern S, Lee D, Wang V, Leysens M, Higgins B *et al.*: **Dysregulated expression of androgen-responsive and nonresponsive genes in the androgen-independent prostate cancer xenograft model CWR22-R1.** *Cancer Res* 2000, **60**:6134-6141.

33. •• Clark EA, Golub TR, Lander ES, Hynes RO: **Genomic analysis of metastasis reveals an essential role for RhoC.** *Nature* 2000, **406**:532-535.
This paper shows the power of parallel gene-expression analysis to identify critical pathways and genes, in this case related to metastasis. RhoC is identified as a critical regulator of tumor-cell invasion in an *in vivo* model, an observation confirmed by expressing RhoC constructs.

34. •• Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP *et al.*: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344**:539-548.
The authors investigate hereditary breast cancers arising in patients with *BRCA1* or *BRCA2* mutations to determine whether they have a characteristic gene-expression profile in comparison to sporadic tumors. This proves to be the case. *BRCA1* tumors cluster together particularly strongly. A sporadic case that clusters in this group has methylation of the *BRCA1* promoter and low expression of *BRCA1*. In addition to providing clues to the pathogenesis of hereditary breast cancer, these observations raise the possibility that gene-expression profiling may be used to classify tissue samples in complex diseases to aid gene discovery.

35. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29**:137-140.

36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.

37. Shatkay H, Edwards S, Wilbur WJ, Boguski M: **Genes, themes and microarrays: using information retrieval for large- scale gene analysis.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:317-328.

38. Stapley BJ, Benoit G: Biobibliometrics: **information retrieval and visualization from co-occurrences of gene names in Medline abstracts.** *Pac Symp Biocomput* 2000:529-540.

39. Jenssen TK, Vinterbo S: **A set-covering approach to specific search for literature about human genes.** *Proc AMIA Symp* 2000:384-388.

40. Seungchan K, Dougherty EK, Chen Y, Krishnamoorthy S, Meltzer P, Trent JM, Bittner M: **Multivariate measurement of gene expression relationships.** *Genomics* 2000, **67**:201-209.

41. Akutsu T, Miyano S, Kuhara S: **Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function.** *J Comput Biol* 2000, **7**:331-343.

42. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**:601-620.

43. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P: **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.** *GenomeBiology.com* 2000, 1.

44. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.

45. •• Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409**:533-538.
This study is an important step towards using microarrays to map regulatory elements in yeast. The authors take advantage of the complete sequence to

construct arrays that allow them to identify 200 putative new targets for the transcription factors SBF and MBF.

46. •• Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al.*: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**:2306-2309.

As in Iyer *et al.* [45••], the authors use microarrays to identify yeast genes directly regulated by two transcription factors, Gal4 and Ste12. Expression arrays lend themselves to identifying patterns of co-regulation, and these two studies take the next step of linking this information to genomic sequence.

47. Livesey FJ, Furukawa T, Steffen MA, Church GM, Cepko CL: **Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx.** *Curr Biol* 2000, **10**:301-310.

48. •• Bruder CE, Hirvela C, Tapia-Paez I, Fransson I, Segraves R, Hamilton G, Zhang XX, Evans DG, Wallace AJ, Baser ME *et al.*: **High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH.** *Hum Mol Genet* 2001, **10**:271-282.

This study shows the outstanding quantitative data that can be obtained with array-format comparative genomic hybridization when large-insert genomic clones are arrayed. The authors are able to map constitutional deletions in the *NF2* locus with high resolution and precision. The potential of this approach for disease gene discovery is evident.

49. •• Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-126.

Although this study involves expression data from yeast, it provides a valuable model for the integration of data from 300 conditions. Using this database, the authors can place eight anonymous genes into distinct functional pathways.

50. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G *et al.*: **Experimental annotation of the human genome using microarray technology.** *Nature* 2001, **409**:922-927.